Evaluating Apertus Models for Multilingual Legal Summarisation of Swiss Supreme Court Decisions

Kaushik Karthikeyan
Department of Computer Science
ETH Zurich
Zurich, Switzerland

Jonas Mirlach
Department of Computer Science
ETH Zurich
Zurich, Switzerland

Max Neuwinger
Department of Computer Science
ETH Zurich
Zurich, Switzerland

***Abstract*—Headnotes are concise summaries of court decisions that distill key legal points, enabling legal practitioners to navigate complex rulings efficiently. In this work, we evaluate the performance of the open-source, Swiss-focused Apertus family of models for the task of Swiss landmark decision summarisation. Our experiments establish the fully fine-tuned Apertus 8B model as the top performer within the Apertus family. It sets a robust open-weight baseline that is competitive with leading proprietary models on key lexical similarity metrics. However, we identify a significant adaptation challenge: while fine-tuning successfully teaches the models to adopt the specific structural format of Swiss headnotes, they often struggle to maintain the underlying legal reasoning. This results in a performance "regret" where models prioritize superficial stylistic alignment over logical coherence, particularly in cross-lingual settings. Our code and fine-tuned checkpoints are publicly available[1].**

***Index Terms*—LLM, Apertus, LoRA finetuning, Legal NLP**

I. Introduction

Landmark decisions of the Swiss Federal Supreme Court are a core reference point for legal reasoning. Practitioners commonly approach them through short headnotes or maxims rather than by reading the full text [1]. Producing these headnotes is labor intensive and depends on scarce expert time from judges and clerks. At the same time, large language models (LLMs) are increasingly used as assistants in legal workflows, also for drafting summaries of long court decisions [2]. However, state of the art performance in legal summarization is still largely achieved by proprietary models, while fine tuned open weight systems have so far struggled to match their quality on demanding legal tasks [1]. In this work, we therefore turn to Apertus [3], an open-weight multilingual model whose training already concentrates on German, French and Italian material that is highly relevant for Switzerland. We hypothesize that this combination of language and domain alignment with fine tuning may enable Apertus to close or even invert the quality gap to proprietary systems on Swiss judicial texts. This leads to the research question:

To what extent does the choice between parameter-efficient adaptation (LoRA) and full fine-tuning determine the ability of a multilingual open-weight model to reach or exceed the performance of proprietary LLMs in Swiss legal summarization?

We address this by making the following contributions:

1) We provide a comparative study of adaptation strategies for the Apertus model family (8B and 70B), demonstrating that while LoRA succeeds in capturing structural legal formatting, it can lead to a "performance regret" in reasoning quality that is only resolved through full fine-tuning.
2) We show that a fully fine-tuned Apertus 8B model establishes a new open-weight baseline for Swiss legal NLP, outperforming frontier models like GPT-4o and Claude 3.5 Sonnet in lexical overlap metrics (BERTScore and ROUGE-L).
3) We identify and categorize specific failure modes—including "mode collapse" in LoRA variants and "topical hallucinations" in FFT variants—providing a roadmap for more robust domain-specific legal LLM alignment.

II. Background

A. Swiss Landmark Decisions Summarisation

The Swiss Landmark Decisions Summarisations (SLDS) dataset[2] comprises over 20,000 landmark decisions issued by the Swiss Federal Supreme Court between 1954 and 2024. The decisions are written in German, French, or Italian, and each decision is accompanied by professionally authored headnotes summarising the ruling in all three languages. This results in approximately 60,000 decision–headnote pairs, allowing both monolingual and cross-lingual evaluation of legal summarisation systems [1].

B. Apertus

Apertus [3] is a fully open multilingual large language model suite released in 2025 under the Swiss AI Initiative, coled by EPFL and ETH Zurich, with support from the Swiss National Supercomputing Centre (CSCS). The release comprises two decoder-only Transformer models at 8B and 70B parameters. Both models are pretrained on 15 trillion tokens, with roughly 40% of the pretraining data being non-English, and the project explicitly positions the suite as strong in multilingual settings, including Switzerland's national and regional languages. In addition, Apertus reports targeted resources for

[1]Available at https://github.com/kauuu/apertus-finetuning-lsai and https://hf.co/collections/kkaushik02/apertus-finetuned-on-slds respectively

[2]The dataset is available on HuggingFace.

underrepresented varieties (e.g., Romansh instruction data and Swiss-German dialect instructions), which motivates its evaluation for Swiss legal NLP tasks involving German, French, and Italian court decisions.

C. LoRA

Low-Rank Adaptation (LoRA) [4] is a parameter-efficient fine-tuning method in which only a small number of trainable low-rank adapter matrices are introduced into a pre-trained model, while the original model parameters remain frozen. This substantially reduces the number of parameters that need to be trained and lowers the computational and memory requirements of fine-tuning. Moreover, LoRA enables multiple task-specific adapters to be trained and deployed on top of a shared base model, facilitating efficient model reuse across tasks. These properties make LoRA a particularly attractive approach for fine-tuning large language models in resource-constrained and multi-task settings.

III. Experimental Setup

Our experimental pipeline consists of two primary stages: fine-tuning the Apertus models and evaluating their performance on the SLDS dataset. We leveraged the Apertus fine-tuning recipes, adapting the hyperparameters to suit the specific scale of the 8B and 70B architectures. Our configuration choices were specifically informed by recent empirical scaling laws established for low-rank adaptation in production-scale models [5], [6].

A. Fine-tuning and Optimization

We fine-tuned the Apertus models using both Full Fine-Tuning (FFT) and Parameter-Efficient Fine-Tuning (PEFT) via LoRA. For all training regimes, we utilized the `SFTTrainer` from the TRL library and implemented a completion-only loss masking strategy to ensure optimization was driven exclusively by the generation of the legal headnote.

The optimization parameters were differentiated by model scale to maintain numerical stability while attempting to reach the "low-regret" regime described by Schulman [5], where LoRA performance matches FFT. For the 8B variant, we employed a peak learning rate of 2×10^{-4} across both FFT and LoRA runs. While Schulman observes that the optimal LoRA learning rate is typically $10\times$ higher than that of FFT, our preliminary sweeps indicated that maintaining parity at 2×10^{-4} provided the most stable convergence for the Apertus architecture on judicial texts. Both 8B configurations utilized a cosine scheduler decaying to a 10% floor over two epochs.

The 70B variant required a more conservative optimization strategy. For LoRA, we utilized a peak learning rate of 5×10^{-5} over two epochs. The FFT variant was trained for a single epoch with a peak learning rate of 1×10^{-5} and a 5% minimum learning rate floor. This reduction in learning rate and duration for the larger model was necessary to mitigate catastrophic forgetting and manage the increased gradient instability often encountered at the 70B parameter scale.

B. LoRA Configuration and Capacity

Our implementation of LoRA follows the standard parametrization where the update is scaled by α/r. Based on the findings of Schulman [5], we applied LoRA to all linear layers, including both the attention and MLP blocks. This is critical because the MLP layers house the majority of a model's knowledge capacity; attention-only LoRA has been shown to underperform significantly, even when matching parameter counts through higher ranks.

Regarding the choice of rank, Kirkby and O'Neill [6] identify a capacity saturation point where low-rank adapters begin to lag behind FFT as dataset size increases. Given that each legal decision in the SLDS corpus is high-context (often exceeding 5,000 tokens), we selected a rank $r = 32$ to ensure sufficient expressivity. While Kirkby and O'Neill suggest that rank 8 or 16 is often sufficient for simpler production tasks, the linguistic complexity of Swiss multilingual law suggests a higher-rank manifold is required to avoid the logarithmic decay of performance relative to rank.

The scaling factor α was set to 128 for the 8B model and 64 for the 70B model. Following Schulman's analysis of parametrization invariances, we maintained the standard initialization where matrix B is zero-initialized. This creates an implicit schedule where the effective learning rate increases as training progresses and the B matrix grows in spectral norm relative to the A matrix.

C. Evaluation

We conduct our evaluation in line with the SLDS evaluation framework[3] to ensure direct comparability with previously reported results. The framework evaluates summarisation quality on the SLDS test set using a combination of standard automatic metrics and LLM-based judgment. Specifically, we report BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore to measure lexical overlap, n-gram consistency, and semantic similarity between generated summaries and the reference headnotes. In addition, we employ an LLM-as-a-Judge setup, following [1], where a LLM assigns quality scores to summaries based on correctness, completeness, and faithfulness with respect to the source decision. Although we use the same judge model (**DeepSeek-V3**) and prompts as in [1], we observe systematic differences in absolute judge scores compared to those originally reported. To enable a fair comparison of our fine-tuned Apertus models, we therefore reran the evaluations for Phi-3.5, Llama 3.2, and Qwen2.5 models from [1] using our evaluation setup. Due to cost and availability constraints, we did not rerun the evaluations for GPT-4o, Claude 3.5 Sonnet, DeepSeek-R1, and o3-mini; their results are included as reported in [1] for reference.

IV. Results and Discussion

The evaluation results on the SLDS test set, summarized in Table I, reveal a nuanced landscape of performance across proprietary and open-weight models.

[3]Available at https://github.com/rolshoven/slds-eval

Model	Variant	BERTScore ↑	BLEU ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	JUDGE ↑
Phi-3.5-mini	fine-tuned (LoRA)	19.76 ± 2.20	11.76 ± 0.03	33.18 ± 1.51	12.42 ± 0.90	20.37 ± 0.94	18.12 ± 2.75
Llama 3.2 3B	fine-tuned (LoRA)	8.57 ± 4.09	10.65 ± 0.03	28.35 ± 2.06	11.80 ± 1.16	19.32 ± 1.32	18.97 ± 3.01
Qwen2.5 0.5B	fine-tuned (LoRA)	2.45 ± 3.83	9.28 ± 0.03	24.89 ± 1.70	9.54 ± 1.05	17.40 ± 1.08	8.46 ± 2.42
Qwen2.5 1.5B	fine-tuned (LoRA)	22.94 ± 1.80	14.02 ± 0.03	34.94 ± 1.53	13.71 ± 1.01	22.19 ± 0.88	17.35 ± 2.71
Qwen2.5 3B	fine-tuned (LoRA)	26.40 ± 2.24	14.52 ± 0.04	36.62 ± 1.68	15.67 ± 1.31	24.35 ± 1.16	25.13 ± 2.71
Qwen2.5 7B	fine-tuned (LoRA)	29.19 ± 1.96	16.23 ± 0.04	38.76 ± 1.63	17.59 ± 1.30	25.68 ± 1.17	30.71 ± 2.93
Qwen2.5 14B	fine-tuned (LoRA)	30.81 ± 2.41	16.53 ± 0.04	40.01 ± 1.89	18.82 ± 1.47	26.86 ± 1.38	35.59 ± 3.06
Apertus 8B	zero-shot	5.25 ± 1.79	5.69 ± 0.02	27.57 ± 1.53	9.28 ± 0.80	15.19 ± 0.77	27.19 ± 2.49
Apertus 8B	one-shot	16.64 ± 2.11	9.77 ± 0.03	32.63 ± 1.73	12.64 ± 1.07	19.43 ± 0.98	27.59 ± 2.37
Apertus 8B	fine-tuned (LoRA)	24.99 ± 2.14	12.30 ± 0.04	34.81 ± 1.65	15.33 ± 1.19	23.13 ± 1.15	24.64 ± 2.47
Apertus 8B	fine-tuned (full)	**31.80 ± 2.22**	16.29 ± 0.04	40.50 ± 1.88	**19.73 ± 1.55**	27.40 ± 1.37	34.13 ± 2.51
Apertus 70B	zero-shot	9.74 ± 1.94	6.91 ± 0.02	30.48 ± 1.72	11.69 ± 0.92	16.25 ± 0.81	34.83 ± 2.31
Apertus 70B	one-shot	14.00 ± 2.61	8.46 ± 0.03	32.91 ± 1.85	13.16 ± 1.05	18.88 ± 1.02	34.22 ± 3.10
Apertus 70B	fine-tuned (LoRA)	3.18 ± 4.16	6.70 ± 0.03	24.44 ± 2.12	9.70 ± 1.28	15.46 ± 1.37	21.49 ± 3.76
Apertus 70B	fine-tuned (full)	6.04 ± 7.42	9.60 ± 0.05	28.94 ± 2.83	14.23 ± 1.71	20.09 ± 1.98	25.52 ± 3.58
GPT-4o	one-shot	30.44 ± 1.74	**31.89 ± 0.25**	**42.12 ± 1.79**	18.92 ± 1.22	25.92 ± 1.05	39.70 ± 2.66*
Claude 3.5 Sonnet	one-shot	5.53 ± 2.00	21.88 ± 0.25	41.86 ± 1.64	19.23 ± 1.19	**27.67 ± 1.20**	41.25 ± 2.90*
DeepSeek-R1	one-shot	20.28 ± 1.45	22.37 ± 0.18	38.30 ± 1.82	15.97 ± 0.85	21.03 ± 0.84	**42.28 ± 2.21***
o3-mini	one-shot	14.18 ± 1.31	20.55 ± 0.17	34.77 ± 1.43	11.92 ± 0.69	18.21 ± 0.67	34.82 ± 2.41*

*obtained from a separate JUDGE run

TABLE I

PERFORMANCE COMPARISON ON THE SLDS TEST SET. BERTSCORE F1 IS RESCALED USING A BASELINE, WHILE BLEU, ROUGE, AND JUDGE SCORES (COMPUTED USING DEEPSEEK-V3 AND SCALED TO 0–100) ARE REPORTED. RESULTS FOR GPT-4O, CLAUDE 3.5 SONNET, DEEPSEEK-R1, AND O3-MINI ARE TAKEN FROM [1]. **BOLD** INDICATES THE BEST OVERALL SCORE; UNDERLINED INDICATES THE BEST SCORE WITHIN EACH CATEGORY.

A. Quantitative Performance

Our primary finding is that the Apertus 8B (Full Fine-Tuned) achieves comparable performance when considering the JUDGE score, effectively establishing a competitive open-weight baseline for Swiss legal tasks. Notably, it achieves state-of-the-art performance in lexical similarity metrics, yielding the highest BERTScore (31.80) and ROUGE-2 (19.73) across all evaluated models, including proprietary ones. This suggests that the strong BERTScore and ROUGE performance is driven by the model's ability to internalise the structural conventions of SLDS headnotes during the fine-tuning phase. By learning the characteristic formatting, the model produces outputs that closely match the expected structure, resulting in higher overlap-based evaluation scores.

However, a significant gap remains in generative quality. While the Apertus models excel at lexical overlap, proprietary models—specifically DeepSeek-R1 (42.28) and Claude 3.5 Sonnet (41.25)—maintain a lead in the JUDGE score. This indicates that while the fine-tuned Apertus models are highly effective at utilizing domain-specific legal jargon, proprietary frontier models still possess a superior capability for synthesizing coherent and logically consistent headnotes.

B. The LoRA Performance Regression

A surprising observation in our results is the consistent performance regression in the LoRA-tuned variants. For both the 8B and 70B architectures, LoRA fine-tuning resulted in lower JUDGE scores compared to their respective one-shot or zero-shot baselines. We find this behavior particularly noteworthy because it persisted across multiple experimental setups; we conducted a hyperparameter search varying the number of training epochs and testing various learning rates, yet the performance dip remained a constant factor.

This "LoRA Regret" appears to stem from a parameter bottleneck that forces the model to prioritize structural mimicry at the expense of deeper reasoning and instruction following. The instability was most pronounced in the 70B LoRA variant (JUDGE 21.49), where the low-rank updates proved insufficient to re-align the model's massive general knowledge base with the rigid constraints of judicial headnotes. This unexpected outcome served as the primary motivation for the qualitative analysis presented in V, as we sought to identify the specific failure modes, such as mode collapse, that were not fully captured by automated lexical metrics.

C. Cross-lingual Robustness

The cross-lingual evaluation, visualized by the heatmaps in Figure 2, reveals a key strength of the Apertus 8B (FFT) model: it maintains high summarization consistency across all nine language pairs in our Swiss-German-French-Italian test set.

Unlike other open-source models evaluated on similar multilingual legal tasks [1], which often exhibit significant performance degradation when the input (decision) and output (headnote) languages differ, the Apertus 8B (FFT) model demonstrates robust cross-lingual transfer. We attribute this robustness primarily to the multilingual pre-training corpus of the base Apertus model.

However, this robustness has a notable limitation. While the model preserves semantic content across languages, our manual analysis indicates it often fails to strictly adhere to the language specified in the instruction prompt (e.g., generating a headnote in Italian when French was requested). This failure mode is not captured by our primary automated evaluation (see Section A for criteria), as the LLM-as-a-Judge was not explicitly prompted to penalize language instruction violations.

This highlights a distinction between the model's inherent multilingual capability and its instruction-following precision in a cross-lingual setting.

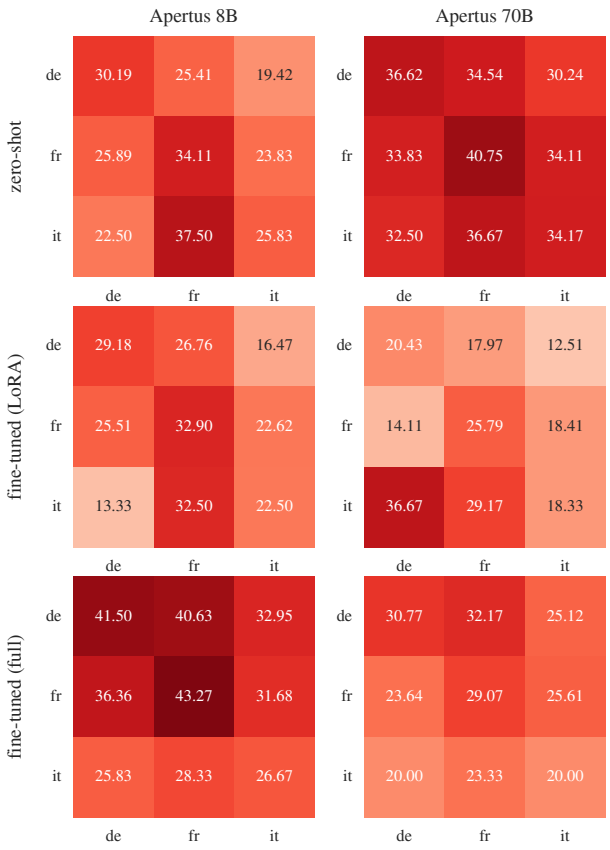

Fig. 1. Cross-lingual SLDS heatmaps (JUDGE score using DeepSeek-V3) for Apertus 8B/70B across training variants. In each panel, x-axis = headnote language and y-axis = decision language.

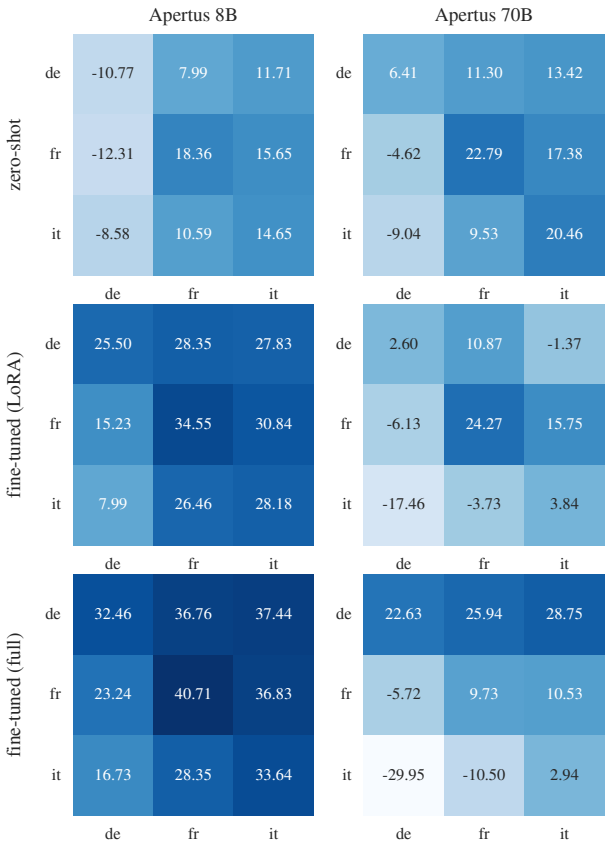

Fig. 2. Cross-lingual SLDS heatmaps (BERTScore-F) for Apertus 8B/70B across training variants. In each panel, x-axis = headnote language and y-axis = decision language.

V. Qualitative Analysis and Discussion

To investigate the performance variance observed in Table I we conducted a manual inspection of sampled headnotes across each language task. This analysis reveals a distinct evolution of model behavior as the adaptation strategy shifts from zero-shot baselines to full fine-tuning. We have included examples in Appendix A.

A. Baseline Behavior: Context Overload

In both zero-shot and one-shot settings, the baseline Apertus 8B and 70B models frequently reverted to a text-completion mode. Rather than generating a concise headnote, the models tended to extend the narrative of the court decision. This behavior likely stems from the extreme length of the input documents; the models' attention mechanisms appear to prioritize the dense context of the legal decision over the specific instructions provided in the prompt. This suggests that future work should explore alternative prompt architectures, such as appending instructions at the end of the context, to maintain task focus.

B. The LoRA Variant: Structural Over-fitting and Mode Collapse

The LoRA fine-tuned models exhibited a paradoxical decline in JUDGE scores despite their training. While these variants successfully adopted the headnote structure seen in the training data — utilizing the characteristic semicolon-delimited and article-first layouts — they suffered from a loss in generation quality.

Moreover, like the Base Models, they occasionally ceased summarization entirely in favor of verbatim extraction of the *Sachverhalt* (Facts) section. By confusing summarization with document completion, the model failed to distill legal meaning. This was accompanied by a distributional bias toward common statutes; the model frequently hallucinated unrelated articles from the SchKG (Debt Enforcement) in cases concerning military tax or contracts, effectively "faking" legal expertise through the injection of high-frequency terms. In extreme cases, the LoRA adapter ($r = 32$) appeared to over-fit on metadata headers, leading to infinite loops of year sequences (e.g., 2024 to 1959) at the expense of semantic processing.

We also occasionally observed "language bleeding", particularly German terms bleeding into the outputs of other languages. This linguistic drift is likely a consequence of the dataset's imbalance, as over 60% of the training decisions are

in German.

C. Full Fine-Tuning: Professional Specialization

Full fine-tuning (FFT) proved necessary to reconcile the rigid structural constraints of Swiss law with the reading comprehension required for accurate summarization. The FFT model consistently achieved high structural fidelity, producing the expected sequence of leading articles, keywords, and concise legal dicta.

Crucially, the linguistic isolation in the FFT variant is significantly improved. The German leakage prevalent in the baseline and LoRA variants was largely resolved, with the model maintaining target-language proficiency even in complex cross-lingual settings. Despite these gains, some limitations remain. The model, influenced by the imbalanced distribution of legal topics in the training data, develops strong priors for frequently encountered laws (e.g., anti-terrorism statutes). During inference, these priors can dominate, causing the model to default to generating a "stereotyped" headnote for the common topic—even when the input case concerns a completely different matter.

D. The Parameter Bottleneck

These technical results suggest a clear parameter bottleneck in the LoRA approach for the legal domain. The limited update capacity appears to force the model to prioritize high-reward formatting patterns over the nuanced reasoning required for judicial summarization. Full fine-tuning bypasses this limitation, allowing the model to integrate domain-specific style and logic into its core weights rather than treating the task as a superficial formatting layer.

VI. Limitations

This study acknowledges several limitations. First, the LLM-as-a-Judge framework relies on DeepSeek-V3; while previous research suggests a high correlation with human experts, the model may still exhibit systemic biases, particularly regarding specific formatting styles. Second, our manual inspection was constrained by a lack of domain-specific legal expertise. Furthermore, as the evaluators did not possess native proficiency in all target languages, they were reliant on machine translation tools, which may obscure subtle nuances. Moreover, due to time-constraints, more training settings could not be tried, which could have potentially improved model performance (for example, lower learning rate for Apertus-70B to avoid catastrophic forgetting).

VII. Conclusion

In this work, we evaluated the Apertus model family for the task of multilingual Swiss legal summarization. Our results demonstrate that Full Fine-Tuning of the Apertus 8B model outperforms leading proprietary models (including GPT-4o and Claude 3.5 Sonnet) in lexical similarity metrics, such as BERTScore and ROUGE.

Crucially, our analysis highlights a trade-off between parameter-efficient fine-tuning (LoRA) and full fine-tuning. We found that LoRA-based models often suffer from mode collapse—prioritizing legal formatting while degrading the actual reasoning of the summary. For high-stakes legal applications in Switzerland, full fine-tuning remains the superior pathway for creating specialized assistants.

Although proprietary models maintain an edge in holistic legal judgment, our work establishes that open-weight, sovereign models like Apertus are highly competitive when properly adapted. The fully fine-tuned Apertus 8B model provides a performant and controllable alternative, setting a strong baseline for open-source legal AI in Switzerland.

Future research must therefore prioritize enhancing instruction-following robustness—specifically to guarantee strict adherence to output language and to eliminate the semantic drifts and catastrophic hallucinations observed in this study—thereby closing the reliability gap for real-world deployment.

References

[1] L. Rolshoven, V. Rasiah, S. B. Bose, S. Hostettler, L. Burkhalter, M. Stürmer, and J. Niklaus, "Unlocking legal knowledge: A multilingual dataset for judicial summarization in Switzerland," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 15 382–15 411. [Online]. Available: https://aclanthology.org/2025.findings-emnlp.832/

[2] D. Jain, M. D. Borah, and A. Biswas, "Summarization of legal documents: Where are we now and the way forward," *Computer Science Review*, vol. 40, p. 100388, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013721000289

[3] P. Apertus, A. Hernández-Cano, A. Hägele, A. H. Huang, A. Romanou, A.-J. Solergibert, B. Pasztor, B. Messmer, D. Garbaya, E. F. Ďurech, I. Hakimi, J. G. Giraldo, M. Ismayilzada, N. Foroutan, S. Moalla, T. Chen, V. Sabolčec, Y. Xu, M. Aerni, B. AlKhamissi, I. A. Mariñas, M. H. Amani, M. Ansaripour, I. Badanin, H. Benoit, E. Boros, N. Browning, F. Bösch, M. Böther, N. Canova, C. Challier, C. Charmillot, J. Coles, J. Deriu, A. Devos, L. Drescher, D. Dzenhaliou, M. Ehrmann, D. Fan, S. Fan, S. Gao, M. Gila, M. Grandury, D. Hashemi, A. Hoyle, J. Jiang, M. Klein, A. Kucharavy, A. Kucherenko, F. Lübeck, R. Machacek, T. Manitaras, A. Marfurt, K. Matoba, S. Matrenok, H. Mendonça, F. R. Mohamed, S. Montariol, L. Mouchel, S. Najem-Meyer, J. Ni, G. Oliva, M. Pagliardini, E. Palme, A. Panferov, L. Paoletti, M. Passerini, I. Pavlov, A. Poiroux, K. Ponkshe, N. Ranchin, J. Rando, M. Sauser, J. Saydaliev, M. A. Sayfiddinov, M. Schneider, S. Schuppli, M. Scialanga, A. Semenov, K. Shridhar, R. Singhal, A. Sotnikova, A. Sternfeld, A. K. Tarun, P. Teiletche, J. Vamvas, X. Yao, H. Zhao, A. Ilic, A. Klimovic, A. Krause, C. Gulcehre, D. Rosenthal, E. Ash, F. Tramèr, J. VandeVondele, L. Veraldi, M. Rajman, T. Schulthess, T. Hoefler, A. Bosselut, M. Jaggi, and I. Schlag, "Apertus: Democratizing open and compliant llms for global language environments," 2025. [Online]. Available: https://arxiv.org/abs/2509.14233

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[5] J. Schulman and T. M. Lab, "Lora without regret," *Thinking Machines Lab: Connectionism*, 2025, https://thinkingmachines.ai/blog/lora/.

[6] M. Kirkby and C. O'Neill, "Practical LoRA research: Fine-tuning at scale: What LoRA gets right (and LoRA-XS doesn't)," Parsed Research, Oct 2025, accessed: 2025-12-18. [Online]. Available: https://parsed.com/research/practical-lora-research

APPENDIX

LLM-AS-A-JUDGE METRIC

We utilised the DeepSeek-V3 model for LLM-as-a-Judge and followed the same prompt used by [1]. The metrics they used to judge the outputs were:

1) *Accuracy & Faithfulness*: How well does the Model-Generated Headnote match the essential legal meaning and intent of the Official Headnote?
2) *Completeness & Relevance*: Does the Model-Generated Headnote include all important points that the Official Headnote emphasizes, without adding irrelevant details?
3) *Clarity & Coherence*: Is the text well-organized, easy to understand, and coherent in style and structure?
4) *Articles*: Do the same legal articles (prefixed "Art.") appear correctly and completely in the Model-Generated Headnote as in the Official Headnote?
5) *Considerations*: Do the same considerations (prefixed "E." in German or "consid." in French/Italian) appear correctly and completely in the Model-Generated Headnote as in the Official Headnote?

QUALITATIVE ANALYSIS OF MODEL OUTPUTS

In this section, we will list some limitations we observed with the various model variants we studied, along with concrete examples.

Task Misalignment

The base model(s) (and sometimes LoRA variants) struggle to stick to the required output format. The expected format is a concise headnote, but the model tends to produce a more detailed summary or continue the landmark decision like a text completion model.

Golden Headnote (Expert Output)

Art. 10 Abs. 2 lit. b, Art. 16 Abs. 2 lit. i und Art. 24 AVIG; Art. 41a Abs. 1 AVIV; Anspruch auf Kompensationszahlungen bei teilweiser Arbeitslosigkeit. Begriff des Zwischenverdienstes bei Ausübung mehrerer Teilzeitbeschäftigungen vor Eintritt der Arbeitslosigkeit; Zusammenfassung der Rechtsprechung (E. 5). Ein Anspruch auf Kompensationszahlungen besteht nur, wenn der Bruttotagesverdienst der versicherten Person geringer ist als das Taggeld, das sie bei Vollarbeitslosigkeit erhielte; Anwendungsfall (E. 6).

Apertus-8B-Base Output (Incorrect Format)

Kopfnotiz zur führenden Entscheidung: In diesem Urteil wird die Frage geklärt, ob ein teilweise arbeitsloser Versicherter Anspruch auf eine Entschädigung aus der Arbeitslosenversicherung hat, wenn er weiterhin einen Teilzeitjob ausübt und dabei mehr verdient als die ihm zustehende Arbeitslosenentschädigung. Das Bundesgericht bestätigt die Rechtsauffassung der Vorinstanz, wonach das Einkommen aus dem fortbestehenden Teilzeitjob als angemessen zu qualifizieren ist, solange es den Betrag der möglichen Arbeitslosenentschädigung nicht übersteigt. Dies steht im Einklang mit der Rechtsprechung zum angemessenen Lohn nach Art. 16 Abs. 2 lit. i AVIG und zur Berechnung der Arbeitslosenentschädigung bei teilweiser Erwerbslosigkeit. Die Argumentation der Beschwerdeführerin, dass ihr gesamtes Einkommen aus beiden Teilzeitjobs zusammenzuzählen sei, um den Anspruch auf eine Entschädigung zu beurteilen, wird abgelehnt, da dies nicht der geltenden Rechtsprechung entspricht. Der Beschwerdeführer erhält somit keine Entschädigung, da sein Einkommen aus dem fortbestehenden Teilzeitjob höher ist als die mögliche Arbeitslosenentschädigung.

Topic Drift

A severe failure mode observed in the fully fine-tuned model is catastrophic hallucination: in some instances, it generates a coherent headnote for a completely unrelated legal case. For example, when prompted with a decision on preschool care, the model produced a detailed headnote concerning debt enforcement—a different legal domain entirely. This indicates profound overfitting, where the model recalls and outputs memorized patterns from the training set instead of performing the conditional task of summarizing the provided input.

Golden Headnote (Expert Output)

Art. 9 Cst.; art. 132 al. 2 et 3 de la loi genevoise du 26 septembre 2010 sur l'organisation judiciaire; art. 6 de la loi genevoise du 12 décembre 2019 sur l'accueil préscolaire; art. 13 du règlement de la Ville de Genève pour l'accueil préscolaire; nature juridique (droit public ou privé) d'une clause contractuelle fixant le prix de pension pour l'accueil d'un enfant en âge préscolaire par une association privée exclusivement par renvoi à une règlementation communale. Nature juridique d'un avenant modifiant le prix de pension pour l'accueil d'un enfant en âge préscolaire par une association privée. Détermination du caractère public ou privé des prestations caractéristiques d'une telle relation contractuelle. L'accueil d'enfants en âge préscolaire n'étant pas une tâche publique au sens des dispositions légales applicables, cette prestation ressortit au droit privé. En revanche, la fixation exclusive du prix de pension par la Ville de Genève poursuit un

Apertus-8B-FFT Output (Topic Drift)

Art. 56 LP; art. 37 ORFI; art. 132 al. 2 et 3 LOJ/GE; art. 4 al. 1 LPA/GE; qualification de la communication de l'état des charges. La communication de l'état des charges est un acte de poursuite au sens de l'art. 56 LP (consid. 4.2).

In the below example, the input decision (Art. 135 CP) and the generated headnote (Art. 2 LAQ/SI) belong to distinct, non-overlapping chapters of Swiss law (general criminal code vs. specific anti-terrorism legislation). The model, however, appears to have learned a broad, fuzzy category of "serious crimes involving disseminated media." When triggered, this internal category activates the most statistically salient template—likely a frequently seen terrorism case—leading it to generate a headnote for the wrong legal domain entirely.

Golden Headnote (Expert Output)

Art. 135 CP; rappresentazione di atti di cruda violenza. Elementi costitutivi del reato (consid. 4.1).

Apertus-8B-FFT Output (Topic Drift)

Art. 2 cpv. 1 della legge federale che vieta i gruppi "Al-Qaïda" e "Stato islamico" nonché le organizzazioni associate (LAQ/SI); reato di sostegno a organizzazioni terroristiche; nozione di propaganda; dolo eventuale. Nozione di propaganda (consid. 5.2.2). Azioni propagandistiche organizzate (consid. 5.2.2). Clausola generale di promozione in altro modo delle attività dei gruppi o delle organizzazioni vietati (consid. 5.2.3). Nozione di propaganda (consid. 5.2.2). Clausola generale di promozione in altro modo delle attività dei gruppi o delle organizzazioni vietate (consid. 5.2.3). La LAQ/SI può essere commessa con dolo eventuale (consid. 5.7.2). Condividendo un filmato di propaganda dello "Stato islamico", l'insorgente ha accettato il rischio di promuovere in altro modo le attività dell'organizzazione vietata e di rafforzare il potenziale dell'organizzazione. Ha quindi agito con dolo eventuale (consid. 5.7.3).

Degenerate Loops

In some instances, the base and LoRA models demonstrate a tendency to get stuck in a degenerate loop of listing terms in the same sentence:

Apertus-8B-Base Output (Degenerate Loop)

Art. 31 Abs. 1bis, 8c Abs. 2, 47, 23 und 76a BV; Art. 77 Abs. 2, 80 und 89 BPR; Art. 37 Abs. 2 VZG; Art. 31 Abs. 1bis BPR; Art. 24 Abs. 3 BPR; Art. 76a BPR; Art. 80 Abs. 1 BPR; Art. 77 Abs. 2 BPR; Art. 31 Abs. 1bis BPR; Art. 82 LTF; Art. 89 BPR; Art. 77 Abs. 2 BPR; Unterlistenverbindungen zwischen Parteien verschiedener politischer Ausrichtung; Begriff der Gruppierung; Gleichheit der Wahlchancen; Wahlfreiheit; Wahlrecht; demokratische Rechte; rechtliche Grundlage; Gesetzesänderung; Parlamentsdebatten; Praxisänderung; Rechtsanwendung; Verwaltungsrichtlinien; Richtlinien des Bundesrates; Rechtsprechung; Verfassungsgrundsätze; Wahlsystem; Wahlverfahren; politische Parteien; Parteistrukturen; Wahlallianzen; Listenverbindungen; Unterlistenverbindungen; Wahlkreis; Wahlkreiseinteilung; Mandatsverteilung; Wahlresultat; Stimmengewichtung; Sitzverteilung; Wahlquorum; natürliches Wahlquorum; Wahlbeteiligung; Stimmabgabe; Wahlentscheid; Wahlausgang; Wahlresultat; Wahlergebnis; Wahlresultate; Wahlanalyse; Wahlstatistik; Wahlforschung; Wahlkampf; Wahlwerbung; Wahlpropaganda; Wahlkampffinanzierung; Wahlkampfkosten; Wahlkampfspenden; Wahlkampfunterstützung; Wahlkampftätigkeit; Wahlkampfak-

tivitäten; Wahlkampfmaßnahmen; Wahlkampfmittel; Wahlkampfstrategien; Wahlkampfziele; Wahlkampfprogramme; Wahlversprechen; Wahlplakate; Wahlwerbespots; Wahlwerbeaktionen; Wahlwerbekampagnen; Wahlwerbung im Internet; Wahlwerbung in sozialen Medien; Wahlwerbung auf Plakaten; Wahlwerbung an öffentlichen Orten; Wahlwerbung in Zeitungen; Wahlwerbung im Radio; Wahlwerbung im Fernsehen; Wahlwerbung in Kino; Wahlwerbung auf Flugblättern;

Apertus-8B-LoRA Output (Degenerate Loop)

Extrait de l'arrêt publié: Erwägungen Extrait des considérants: 3. 3.1 Les conjoints en union conjugale effective et les personnes en union enregistrée effective (cf. art. 9 al. 1 let. b LIFD [RS 642.11]) sont traités sur le plan procédural comme des contribuables "indépendants" (HUNZIKER/MAYER-KNOBEL, in: Loi fédérale du 14 décembre 1990 sur l'impôt fédéral direct, commentaire suisse du droit fiscal [ci-après: Comm. LIFD], 4ème éd. 2022 , nos 3 et 4 ad art. 9 LIFD; sur les conditions d'une séparation légale ou de fait: ATF 2024/2023/2022/2021/2020/2019/2018/2017/2016/2015/2014/2013/2012/2011/2010/2009/2008/2007/2006/2005/2004/2003/2002/2001/2000/1999/ 1998/1997/1996/1995/1994/1993/1992/1991/1990/1989/1988/1987/1986/1985/1984/1983/1982/1981/1980/1979/1978/1977/1976/1975/1974/1973/1972/1971/1970/1969/1968/1967/1966/1965/1964/1963/1962/1961/1960/1959/195

Language Bleeding

The LoRA fine-tuned model exhibits a critical weakness in instruction adherence regarding language. It occasionally generates a single headnote that mixes multiple languages (e.g., German, French, Italian) within the same output,

Apertus-8B-LoRA Output (Language Bleeding)

Art. 5 cpv. 2 LPT; espropriazione materiale. Quando è dato un caso di rifiuto di attribuire un fondo alla zona edificabile, non si è in presenza di un'espropriazione materiale (consid. 3.3.1). Espropriazione materiale. Lorsque l'on se trouve en présence d'un refus d'affectation à la zone à bâtir, on n'est pas en présence d'une expropriation matérielle (consid. 3.3.1). Enteignung materiell. Wenn ein Grundstück nicht zu einer Bauzone eingewiesen wird, liegt keine materielle Enteignung vor (E. 3.3.1).